JOURNAL OF PARALLEL AND DISTRIBUTED COMPUTING 55, 1–31 (1998)
ARTICLE NO. PC981489

Architecture of a Terabit Free-Space Intelligent Optical Backplane

Ted H. Szymanski

Department of Electrical and Computer Engineering, McMaster University, Hamilton, Ontario, Canada L8S 4K1

and

H. Scott Hinton

Department of Electrical and Computer Engineering, University of Colorado at Boulder, Colorado

Received June 22, 1994; revised December 10, 1997; accepted July 27, 1998

Optical technologies can support thousands of high bandwidth optical channels to/from a single CMOS integrated circuit, and can thus allow for the construction of novel bandwidth-intensive computing architectures which are no longer constrained by conventional electronic wiring limitations. In this paper, the architecture of a dynamically reconfigurable *Intelligent Optical Backplane* is described. The backplane consists of a large number of parallel optical channels (typically 1000–10,000 bits) spaced a few hundred micrometers apart. The optical channels are arranged into upstream and downstream rings, where the channel access protocols are implemented by "smart pixel arrays." The architecture exploits the *bandwith advantage* of the optical domain and can be dynamically reconfigured to embed conventional interconnection networks, including multiple busses, rings, and meshes. Unlike all-optical and passive optical systems, the proposed backplane is intelligent and can support communication primitives used in shared memory multiprocessing, including broadcasting, multicasting, acknowledgment, flow and error-control, buffering, shared memory caching, and synchronization. The backplane is also manufacturable using existing optoelectronic technologies. A second generation backplane supporting a distributed shared memory multi-processor is under development.

Key Words: free-space; optics; backplane; dynamic; reconfigurable; smart pixels; embeddings; meshes.

1. INTRODUCTION

As the digital information processing markets continue to evolve, they will demand faster and more intelligent hardware platforms. One popular processing paradigm is the high performance digital backplane interconnecting a large number

0743-7315/98 $25.00

of electrical processing boards. These processing boards can contain the computing nodes of large multiprocessing systems or the switching nodes of large ATM switching fabrics. An optical backplane can be modeled as a collection of several nodes partitioned among N printed circuit boards (PCBs) or multi-chip modules (MCMs), that are interconnected through a large number of optical backplane channels, as shown in Fig. 1. Each PCB or MCM is typically composed of multiple processing elements (PEs) and message processors (MPs) to coordinate communications.

Within the computing community, there is a growing awareness that optics can potentially open up a new frontier of computing machines with massive connectivity not previously possible. This paper explores these issues and describes one approach to exploiting bit-parallel optical technology to provide high bandwidth low latency interconnects, the free-space intelligent optical backplane architecture shown in Fig. 1.

The message processors control access to the Z optical backplane channels $\{C_1, C_2, ..., C_Z\}$. The MPs have access to the backplane channels through both X electrical *injector* and Y *extractor* channels (called "access" channels), labeled $\{I_1, I_2, ..., I_X\}$ and $\{E_1, E_2, ..., E_Y\}$, respectively, where typically $X \leqslant Z$ and $Y \leqslant Z$. The injectors provide the capability of injecting data into a selected subset $\mathbb{I}$ of backplane channels, while the extractors are used to extract data from another subset $\mathbb{E}$ of backplane channels. Optical technologies offer a *bandwidth advantage* over electrical technologies, and hence an optoelectronic integrated circuit can support far more optical channels than electrical access channels. The connectivity associated

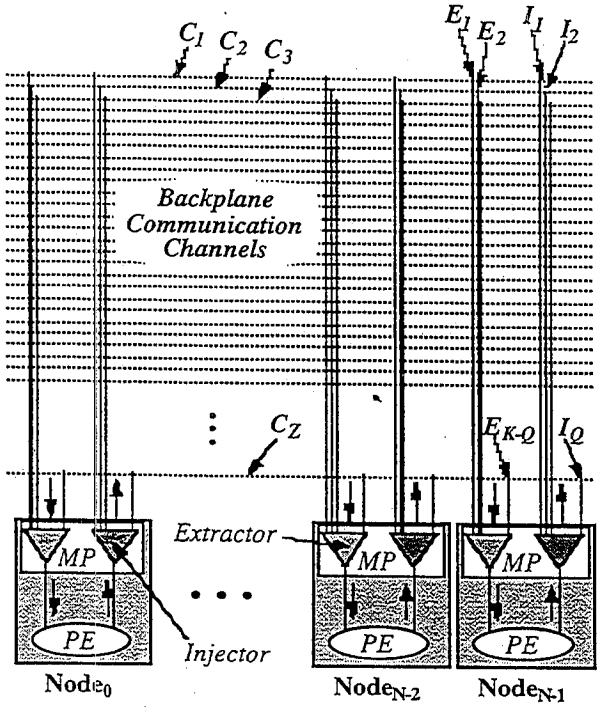

FIG. 1. Backplane connectivity model.

with the combination of both the optical backplane channels and electronic access channels creates a model which can formally be called a reconfigurable "multi-channel array." We will use the phrase *Hyperplane* to denote an intelligent optical backplane based on this multi-channel array model. Provided sufficient optical channels are available, all N node interconnection networks of degree K can be embedded onto the HyperPlane. For example, if every node i has reserved broadcast based channel to all other nodes, i.e., $\mathbb{I} = \{C_i\}$ and $\mathbb{E} = \{C_1, C_2, ..., C_{i-1}, C_{i+1}, ..., C_N\}$, then the topology of the embedded network is equivalent to a conventional "broadcast-and-select" network. The ability to support multiple broadcast channels naturally supports distributed shared memory multiprocessing.

By allowing the MPs to dynamically change the access to the optical backplane channels, a *dynamically reconfigurable* optical backplane is achieved. Each of the optical backplane channels can be reconfigured as a single broadcast channel spanning the entire backplane (end-to-end), or it can be partitioned into multiple smaller channels spanning physically distinct segments of the backplane. Such a programmable backplane could embed a crossbar network at one instant of time, a Butterfly network the next instant, and 2D mesh, hypermesh, or hypercube in yet another instance. The optical backplane can also be dynamically partitioned into multiple subsets, where the subsets can execute independent applications. For example, backplane partitions could embed independent crossbars, butterflies, and meshes simultaneously. This dynamic programmability supports systolic array computing.

Dynamic reconfigurability is important for many reasons. It significantly improves performance by allowing the interconnections between nodes to be dynamically chosen to match the application requirements. It significantly improves fault tolerance by allowing for reconfiguration in the presence of faulty optical channels, which is especially important when using new optical technologies. It improves manufacturability by allowing for the batch fabrication of a single smart pixel array, which can be used in optical backplanes of multiple sizes and diverse applications. In the telecommunications field, the key requirements of digital switch include a low blocking probability for connections and a high bisection bandwidth. In the computing field, the key requirements may include topological compatibility with the existing algorithms (favoring broadcast bus, crossbar, mesh, hypercube, and systolic array interconnections), the ability to reconfigure in the presence of faults, and the ability to reconfigure for higher performance. All these requirements can be met by a single optical backplane which supports dynamic reconfiguration. Finally, with the significant investment of developing a manufacturable optical backplane, it may not be economically viable to create a customized optical interconnect for a small and specific market. A reconfigurable optical backplane can potentially benefit from the economy of scale by appealing to a large established market based upon electrical backplanes.

While there has been a great deal of progress in the development of electrical backplanes, they will ultimately be constrained by the fundamental physical limitations of electronics [22]. Current metal interconnects are limited by the "skin effect" which results in a greater attenuation in transmission lines at high frequencies, and parasitic inductance and capacitance which reduces the usable bandwidth

of the interconnections [22]. Due to packaging constraints, existing PCBs, MCMs, and VLSI ICs are currently limited to typically several hundred electronic I/O pads per substrate [22]. This I/O constraint has heavily influenced multiprocessor interconnections over the years, i.e., see [1, 8]. The power dissipation, parasitic capacitance, and inductance of electronic I/O pads limits the clock rate to typically one GHz per I/O. Together these constraints currently limit the electrical I/O bandwidth of a single substrate to the range of 10–100s Gigabits per second (Gb/s). Advanced new high-speed electrical I/O technologies based upon dynamic equalization [9, 36] do not appear to affect the electrical bandwidth constraint of a single IC significantly (see Section 2). Finally, electronic backplanes will eventually be constrained by the 2D nature of metal traces on a PCB. In contrast, optics currently provides the same degree of massive interconnect as 3D VLSI; using optics, beams of light focused to spots 10s of micrometers wide and spaced 100s of micrometers apart can be routed through 3 D free-space at very high clock rates, without skin effects, with low energy per bit and with no electromagnetic interference. Optics thus represents an attractive technology for bandwidth intensive systems of the future.

One unique feature of the proposed optical backplane is its ability to exploit the *bandwidth advantage* of the optical domain. As described earlier, each IC is currently limited to typically 10–100s Gb/s of electrical I/O bandwidth. However, each IC may have typically 1–10s of Terabits per second (Tb/s) of optical I/O bandwidth. The proposed backplane recognizes and exploits this *bandwidth mismatch*, by allowing a PCB with a limited electrical I/O bandwidth the capability to "tap" a reconfigurable optical interconnect with significantly more optical bandwidth.

The bandwidth advantage can be exploited in three fundamental ways, using space, time and wavelength division multiplexing (i.e., SDM, TDM, and WDM). By exploiting the *spatial parallelism* of optics, the backplane can support more bit-parallel optical channels than electrical channels (also see [2, 25, 31]). By exploiting the temporal advantage of optics, the optical channels can be clocked at much faster rates than the electrical channels (also see [13, 31]). By exploiting the *wavelength parallelism* of optics, the backplane optical channels can use distinct wavelengths. Combinations of these three approaches will likely be used in future systems, resulting in very high aggregate throughputs.

In a TDM HyperPlane, a specific time slot TS_i is equivalent to a single communication channel C_i. Using TDM the optical clock rate must be significantly higher than the electrical clock rate of the processing boards, thus creating many effective parallel channels in the time domain. In a WDM HyperPlane, a specific wavelength l_i is associated with each communication channel C_i. The injectors and extractors for this system could exploit arrays of surface emitting lasers operating over several distinct wavelengths. In a SDM HyperPlane, specific spatial locations are associated with each communication channel C_i. Space division multiplexing exploits the large 2D spatial bandwidth made available on a single die and the large 3D spatial bandwidth made available through free-space.

Figure 2 illustrates one possible organization of the optical backplane, based on the concept of "logically transparent" processing boards (i.e., also see [7, 14, 34] for descriptions of optically transparent technologies). The optical backplane channels

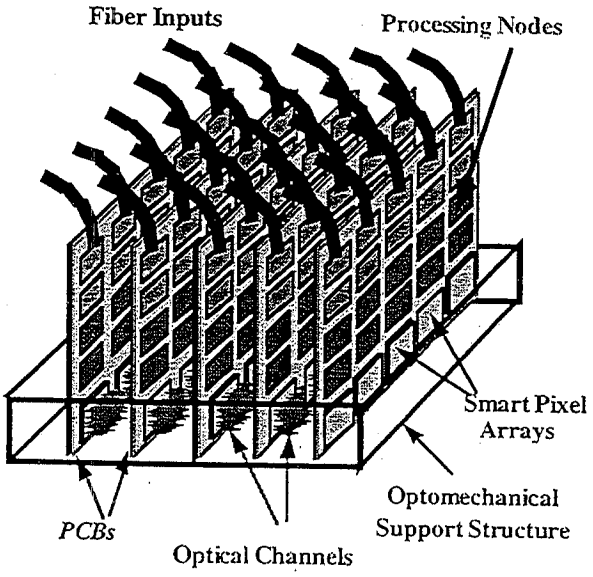

FIG. 2. Free-space optical backplane with SPAs on PCBs.

are created through the use of either imaging optics or 2D microlens arrays between the PCBs, as shown in Fig. 3 (i.e., see [16]). In this architecture, each smart pixel array must have the capability to receive electrical data from the PCB, inject electrical data into the required *optical channels*, monitor all the optical channels for data that need to be extracted, extract data from the optical channels, and deliver the extracted data to the PCB. A unique feature of the proposed backplane is that it is an *intelligent* system. The smart pixel arrays can simultaneously *transport* and *process* terabits of data per second and make decisions on which data to extract according to arbitrary extraction criteria. This unique capability can be used

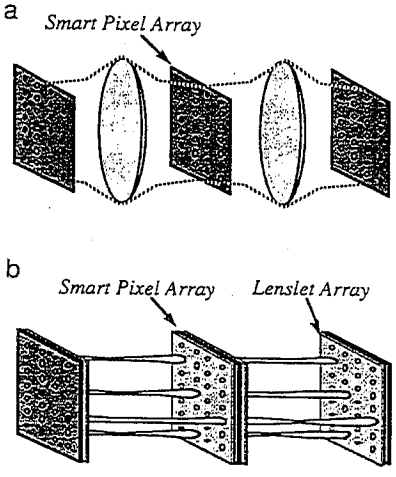

FIG. 3. Two optical imaging technologies: (a) Bulk optics. Imaging-based optical channels. (b) Micro-optics. μ channel-based optical channels.

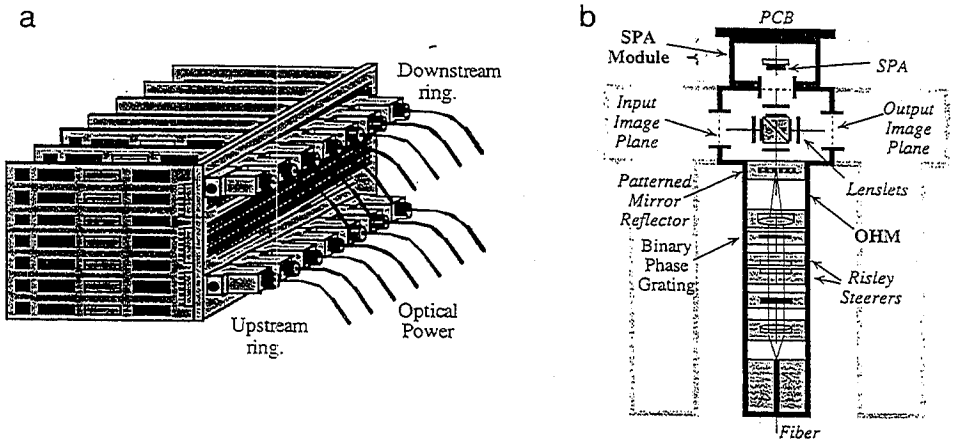

FIG. 4. (a) Opto-electronic backplane with SPAs and OHMs on backplane PCB. (b) Optical hardware module.

to implement communication primitives used in shared memory multiprocessors directly in the optical backplane [31], including point-to-point and multi-point switching, broadcasting, error and flow control, packet acknowledgment, media access control protocols such as token ring, slotted ring, and pipelined bus access schemes, packet buffering, parallel prefix, resource arbitration, snoopy-cache coherence protocols, and synchronization, to name a few.

An alternative organization of the backplane is shown in Fig. 4a, i.e., also see [24, 32]. In this scheme, the smart pixel arrays are mounted on the backplane PCB and optically interconnected with the *Optical Hardware Modules* shown in Fig. 4b. The Optical Hardware Modules accept an incoming array of optical bits and image them onto the SPA. They also provide optical power to the SPA and generate an outgoing array of optical bits. See [11, 24, 26] for detailed descriptions of such modules. The insertable PCBs plug into the backplane using conventional electrical connectors and have access to a subset of the bandwidth of the optical backplane. This scheme isolates the optics to a single static and rigid structure (the backplane). With either of the packaging schemes, each PCB may have between 10 and 100s Gb/s of electrical I/O bandwidth, and the optical backplane may have between 1 and 100s Tb/s of optical bandwidth.

The remainder of this paper includes a detailed discussion of the HyperPlane intelligent optical backplane architecture and the smart pixel arrays required by the architecture. Finally, there will be a discussion of the different networks that can be embedded into the HyperPlane and their performance capabilities.

2. THE HYPERPLANE ARCHITECTURE

An *embedding template* for the *Circular Hyperplane* is shown in Fig. 5. (The Circular HyperPlane includes wrap-around edges.) The template uses a *box* to denote each backplane PCB or MCM. Each PCB has a number of *vertical lines* which

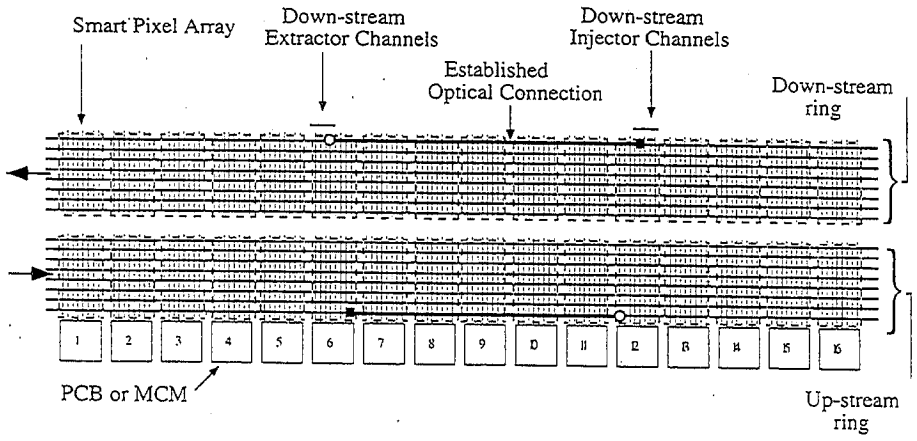

FIG. 5. HyperPlane embedding template.

represent electrical *Injector* and *Extractor* channels to/from the MP. The template has a large number of *horizontal lines* which in this paper represent uncommitted *optical channels*. The phrase *channel* will refer to a collection of parallel bits which are switched together as an indivisible entity and which provide a basic unit of bandwidth. Channels can have any width, such as 32, 64, or 128 bits. The optical backplane will typically support a large number of optical channels, while each smart pixel array will support a smaller number of electrical access channels. The horizontal lines in the 2D template denote optical channels without specifying their precise physical location in 3D free-space. Free-space optics provides a very large 3D spatial parallelism which the 2D template does not reflect.

Boxes, circles, and bold lines represent the connections in the Circular HyperPlane. The solid boxes represent injection points (i.e., connections between electrical injector channels and optical backplane channels), the circles represent extraction points (i.e., connections between optical backplane channels and electrical extractor channels), and the bold lines represent established point-to-point or multi-point optical connections. The number of vertical access channels emanating from a PCB represents the PCB "degree," i.e., a PCB with K vertical access channels can access at most K optical channels. When a connection between PCBs is established, a bold horizontal line is drawn between the optical channel endpoints.

In electrical architectures it is common to construct wide bit-parallel datapaths using a "bit-slice" approach, by operating several narrower datapaths in parallel. The same approach can be used in the HyperPlane; i.e., a 128-bit-wide bus or ring can be achieved by operating several narrower channels (i.e., 32-bit-wide channels) in parallel.

One appealing aspect of the HyperPlane architecture is its immense bandwidth when compared to a conventional electronic system, sufficiently large to embed many other interconnection networks. This is illustrated in the following property (the proof is straightforward). Consider the embedding of a target graph $G(V, E)$ onto the HyperPlane. Using the standard graph-theoretic notation, the graph to be embedded G consists of a set of vertices V and a set of edges E, where each edge

represents a direct connection between precisely two and only two endpoints (i.e., see [5, 19]). The *vertex load* is defined as the maximum number of graph vertices which are embedded into any one HyperPlane node (note that the terms "vertex" and "node" refer to the graph and the HyperPlane, respectively). The *edge load* is defined as the number of graph edges which are embedded into one HyperPlane access channel. Ideally there would be no need for any resource sharing in an embedding; i.e., the vertex load and edge load will be 1.

Property 1. An N-node HyperPlane where every node has K bi-directional access channels can embed any vertex symmetric M-vertex degree-J graph, for $M \geqslant N$ and $J \geqslant K$ with a maximum vertex load of $\lceil M/N \rceil$ and a maximum edge load of $\lceil M/N \rceil \cdot \lceil J/K \rceil$, provided sufficient optical backplane channels can be used.

Traditional graph embedding problems are often NP-complete combinatorial optimization problems where the assignment of graph vertices to host nodes minimizes the edge load and vertex load; i.e., see [2, 3, 5, 19, 23]. The "connection-constrained" nature of the host graph is usually a dominant constraint: there are usually too few host graph edges to perform the embedding in unit edge load, leading to the optimization problem. This problem is greatly alleviated in the HyperPlane, since the free-space optical backplane supports potentially thousands of optical channels, as will be shown in Section 3. In our domain, the dominant constraint is the limited number of electrical channels interfacing to the optical channels.

2.1. Smart Pixel Technology Constraints

This section will demonstrate how *technology constraints* have influenced our proposed optical backplane architecture. A number of technologies can be used to implement smart pixel arrays, including the ELO technology [7], the VCSEL/MSM technology [21], and the "CMOS/SEED" technology [6, 12, 18, 35]. The proposed backplane can use any smart pixel technology, and in the long term the VCSEL technology seems promising. Currently, the CMOS/SEED technology is available for constructing systems, and this technology will be assumed.

The CMOS/SEED technology uses flip-chip bonding to deposit arrays of optical I/O on a conventional silicon CMOS substrate [12, 18]. The process scales well with improvements in the underlying silicon technology; i.e., as the silicon becomes faster the optical processing also becomes faster, and as silicon device density increases the optical I/O density also increases. Projections for the number of optical I/O diodes per CMOS chip are shown in Table 1 [18]. In the year 2001, these devices are expected to contain up to 12,000 optical bits per IC (note that two diodes yield one optical I/O bit using nondifferential signaling, one diode for the input and another for the output). When clocked at the on-chip clock rate of 500 Mhz, each integrated circuit is expected to have an optical I/O bandwidth of up to 6 Tb/s. Hence, a backplane with several optoelectronic devices per PCB (or MCM) will readily scale to 10s of terabits of bisection bandwidth.

It will be difficult for electrical technologies to match these optical I/O bandwidths. Using existing VLSI packaging techniques, the electronic I/O pads are placed around the perimeter of a VLSI die, and these pads are then connected to

TABLE 1

Projected Optical I/O Bandwidth per SPA, Based on [18]

Year	Feature Size (mu)	Gates per die	Optical diodes per die [18]	Gates per optical I/O bit	Optical Clock	Optical I/O BW (Tb/s)
1995	0.35	800 K	6,000	133	200 Mhz	0.6
1998	0.25	2 M	12,000	166	350 Mhz	2.1
2001	0.18	5 M	24,000	208	500 Mhz	6.0
2004	0.12	10 M	40,000	250	700 Mhz	14
2007	0.10	20 M	50,000	400	1 Ghz	25

a mechanical package with electrical pins. The constraint on the number of pads which can be placed around the perimeter of the die and the capacitance and inductance of the mechanical pins and wires limit the electrical I/O bandwidth of a die. Table 2 indicates the Semiconductor Industry Association's (SIA) projections on the number of electrical CMOS I/O pins and the off-chip electrical clock rate per IC. SIA projections are traditionally *conservative*; for example, high-performance microprocessors regularly exceed the SIA projections. In the year 2001, a high-performance IC may contain 2000 electrical pins clocked at 250 Mhz, for an electrical I/O bandwidth of up to 256 Gb/s. These electrical bandwidths are much smaller than the potential optical bandwidths established in the last paragraph, in the neighborhood of 6 Tb/s, thereby illustrating the *bandwidth advantage* of the optical domain. Our proposed smart pixel designs will provide a means to interface this limited electrical bandwidth to the much larger optical bandwidth.

The SIA projections indicate that CMOS I/O pads will be clocked in the Ghz range within a decade. The use of faster I/O pads does not necessarily change the electrical bandwidth constraints of a single IC significantly. While ECL pads can be clocked typically 10 times faster than CMOS, they also have a density approximately 10 times less than CMOS due to their large size, typically 0.5–1 mm^2 per pad. Hence, the aggregate electrical bandwidth of an IC remains approximately the same. The same argument seems to apply to the recently developed Gigabit CMOS I/O technologies based on channel equalization, which also seem to have large I/O pad sizes [9, 36]. Finally, CMOS substrates with VCSEL-based optical I/O will likely be available within a decade [21], and this technology is expected to support

TABLE 2

Projected SPA Electrical I/O Bandwidths, Based on SIA [33]

Year	Off-Chip Elec. Clock (Mhz)	Elec. I/O per die	Electrical I/O BW (Gb/s)	Ratio (Opt. to Ele. BW)
1995	100	900	30	0.60/.030 = 20
1998	175	1350	78	2.10/.078 = 27
2001	250	2000	164	6.00/.164 = 37
2004	350	2600	300	14.0/.300 = 47
2007	500	3600	592	25.0/.592 = 43

optical clock rates in the tens of Gb/s range. An optoelectronic IC with thousands of optical I/O, each supporting several wavelength channels each clocked at 10 s of Gb/s, will have a very large optical I/O bandwidth. Hence, the *bandwidth mismatch* between the electrical and optical domains is expected to continue.

The SIA has conservative projections on gate densities achievable with CMOS technology. The gate density divided by the expected number of optical I/O bits yields the expected number of gates per optical I/O, as shown in Table 1. In the year 2001, a high-performance VLSI die may contain up to 400 gates per optical I/O bit. Hence, a smart pixel array can support a moderate amount of processing on each optical bit, leading to the "intelligence" in our backplane architecture.

In summary, this section has described the range of feasible smart pixel arrays over the next decade, which in turn have motivated our proposed backplane architecture. In the year 2001, a *state-of-the-art* smart pixel array could support 12,000 optical bits and 2000 electrical pins, with optical clock rates of 1 GHz and electrical clock rates of 500 Mhz. Universities usually work with less aggressive technology, and in the year 2001 a *conservative* smart pixel array may contain 1024 optical bits and 512 electronic pins, with an optical clock rate of 500 MHz and an electronic clock rate of 250 Mhz. With these conservative parameters, the optical bandwidth of each SPA is 512 Gb/s and the electrical bandwidth is 64 Gb/s. This design example illustrates the bandwidth advantage of the optical domain (a ratio of 8 to 1).

These bandwidths can be allocated as follows. The system designer may chose the minimum increment of channel bandwidth to be 16 Gb/s. Hence, each conservative SPA will thus support 4 electrical channels at 16 Gb/s each, where each electrical channel is 64 bits wide and clocked at 250 Mhz. Each SPA will also support 32 optical channels at 16 Gb/s each, where each optical channel is 32 bits wide and clocked at 500 Mhz. This design utilizes all of the 64 Gb/s of electrical bandwidth and the 512 Gb/s of optical bandwidth. The SPA will require some straightforward digital logic circuitry at its periphery, i.e., multiplexors and demultiplexors, to convert between the slow wide electrical format and the fast narrow optical format within the die (these will not be explicitly shown). The embedding analysis of Section 3 will illustrate how this bandwidth advantage can be exploited.

2.2. Smart Pixel Design

In this section, the basic design of a smart pixel array which fits within the technology constraints identified in the previous section, and which supports multiple reconfigurable broadcast channels is described. For simplicity, the smart pixel array will use *space-division multiplexing* (SDM) only, although in a real system TDM could be used to increase the optical clock rate, and WDM could be used to increase the width of the datapaths. The organization of a basic smart pixel is shown in Fig. 6. Each pixel consists of optical input and output ports (1 bit each), a programmable delay cell, *concentrator* and *expander* cells, and an address comparator cell. The functions of these cells will be described subsequently.

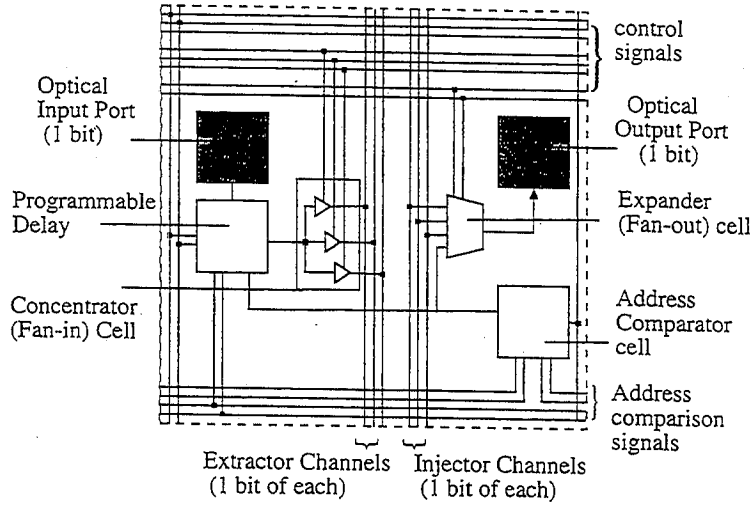

FIG. 6. Basic smart pixel (supporting 3 electrical injector and extractor channels).

Each pixel has four basic states, the *transparent*, *transmitting*, *receiving*, and *transmitting-and-receiving* as shown in Fig. 7. The smart pixels are organized into a 2-dimensional array called a *slice* as shown in Fig. 8. The rows of pixels in the array represent channels which are w bits wide. Multiple communication slices can be implemented on a single VLSI die as shown in Fig. 9. We point out that all these descriptions are representative as many variations of the basic theme can be used.

The optical backplane can be viewed as a distributed crossbar-like switch with multiple I/O ports, one for each PCB, as shown in Fig. 1. Conceptually, one may *slice* up the backplane to isolate the functions and hardware associated with each I/O port. Hence, in our terminology each slice is a self-contained optoelectronic switching module with its own controller circuitry, in essence a slice of a distributed optical backplane. A slice implements the programmable switching between C optical channels, I electronic injector channels, and E electronic extractor channels, where all channels have the same bandwidth. Each slice typically includes an *expander* (or *fan-out*) switching circuit for switching I injector channels onto a subset of $C \geqslant I$ optical channels, and a *concentrator* (or *fan-in*) switching circuit for switching a subset of channels selected from the C optical channels onto E extractor channels. Expanders and concentrators are classic components of computing and communication systems; i.e., see [19]. The VLSI die may contain multiple smaller slices S, where $S \geqslant 1$, each handling a smaller number of channels, or alternatively a single large slice handling all channels (although a larger slice tends to be more complex and have lower fault tolerance). The smart pixel arrays can be designed to support different ratios of electrical to optical I/O bandwidth by adjusting the parameters S, C, I, E, w. Figure 9 illustrates smart pixel arrays with various ratios of optical to electrical bandwidth.

The four basic states of a smart pixel are shown in Fig. 7. By programming the smart pixels appropriately, each optical channel can be configured to span the entire length of the backplane or it can be partitioned into several smaller channel segments. Hence, various topologies can be embedded into the backplane. The optical

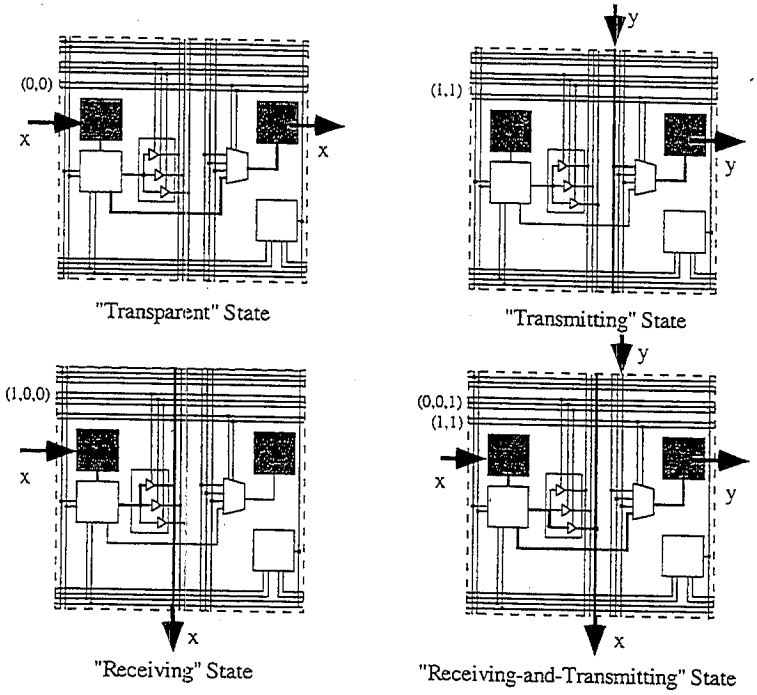

FIG. 7. Four basic states of a Smart Pixel. Active datapaths shown in bold.

backplane can be operated in two general modes, a *reconfigurable* and an *intelligent* mode. In the reconfigurable mode, the backplane can be reconfigured to embed any type of static graph, subject to the constraints on the number of electrical and optical channels. In this mode, the smart pixel arrays do not perform packet processing and dynamic filtering functions. In the intelligent mode, multiple broadcast channels can be embedded into the backplane. The smart pixel arrays process packets of data as they travel down the backplane channels and make decisions dynamically on which packets to extract according to various extraction or filtering criteria.

From the previous section, a conservative SPA design may support 4 electrical channels and 32 optical channels, each with a bandwidth of 16 Gb/s. To achieve a reasonable logic complexity and speed, the conservative 32×32 array can be partitioned into 2 slices, with 16 optical channels and 2 electrical injector and extractor channels per slice. Each slice therefore requires a 16-to-2 concentrator for extracting channels from the backplane, and a 2-to-16 expander for injecting channels onto the backplane. A 16-to-2 concentrator can be made in a regular layout suitable for CMOS VLSI by implementing a one-bit "concentrator cell" within each pixel for extraction, as shown in Figs. 6 and 7. The concentrator cell uses tri-state logic gates to drive an optical bit onto an extractor channel. Similarly, a 2-to-16 expander can be made in a regular layout by implementing a one-bit "expander cell" within each pixel. The expander cell uses 4-to-1 multiplexor to drive an electrical bit onto the optical channel. All of the states in Fig. 7 can be achieved by configuring the states of the concentrator and expander cells in each pixel appropriately.

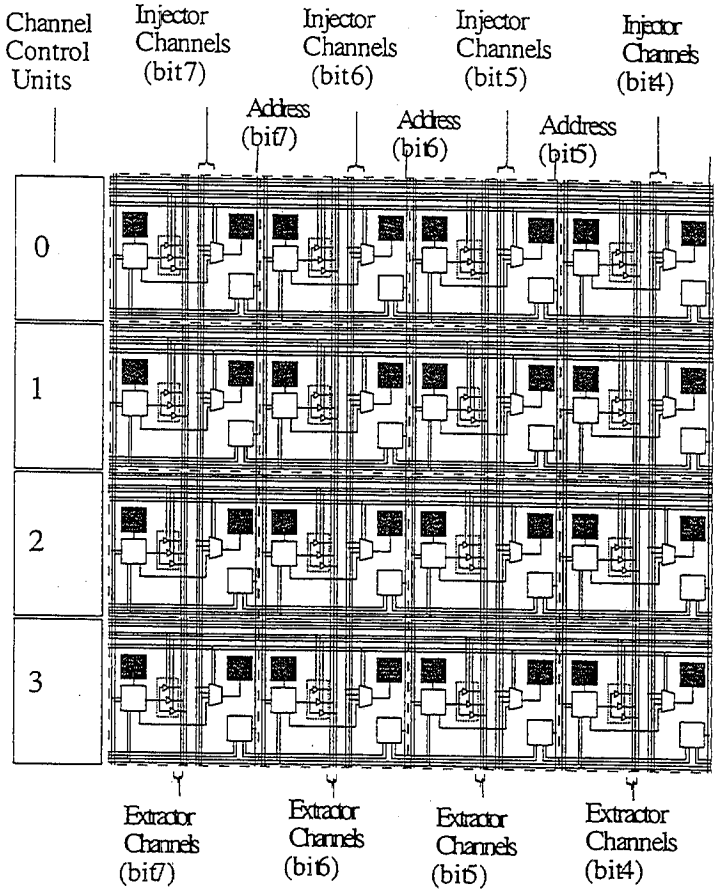

FIG. 8. 2D Slice of pixels with 4 optical channels, 3 injector channels, and 3 extractor channels. (Only the 4 most significant bits of each channel shown.)

In the intelligent mode, each channel processes packet headers looking for packets to extract. Every PCB is assigned a unique address, and the channels look for packets addressed to this PCB. In Fig. 6, the *address comparator cell* in each pixel performs the processing of one bit of the packet header and one bit of the PCB address. In its simplest form, addresses can be *one-hot* encoded, where each PCB is denoted by a "1" in a certain bit position in the header. In this case, the address comparator cell is simple (an AND gate and an OR gate). However, other more complicated addressing schemes can be used to support multicasting or shared memory [31].

With VCSEL technology optical clock rates will approach 10s of Gb/s. The SPAs must have sufficient time to perform address comparisons between the packet headers and the PCB addresses, and if a match occurs to set the state of the concentrators so that the packets can be queued and extracted. To ensure a sufficient amount of time for processing the packet headers, each SPA can contain one or more stages of pipeline latches, so that the optical data resides in a SPA for a few nanoseconds. The programmable delay box in Fig. 6 can be programmed to

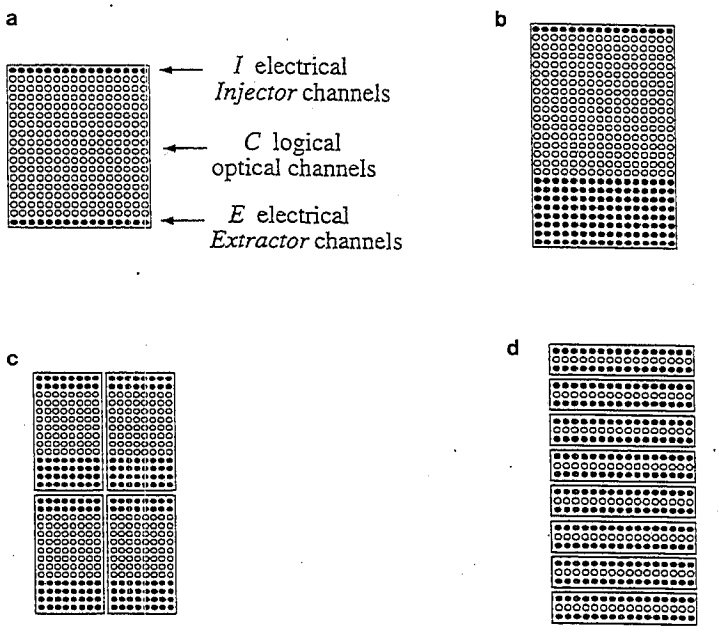

FIG. 9. Smart pixel organizations with varying ratios of optical-to-electrical channels: Ratio Optical/Electrical channels = (a) 16 : 1, (b) 2 : 1, (c) 2 : 1, (d) 1 : 1. (a) $S=1$, $C=16$, $w=16$, $I=1$, $E=1$; (b) $S=1$, $C=16$, $w=16$, $I=1$, $E=8$; (c) $S=4$, $C=8$, $w=8$, $I=2$, $E=4$; (d) $S=8$, $C=1$, $w=16$, $I=1$, $E=1$.

provide several clock cycles of latency, so that the combinational logic can settle before the data is acted on. For example, with an optical clock of 500 Mhz, two pipeline stages would provide 4 ns for the combinational logic to settle. Our experience indicates that the control circuits for the concentrators tend to be among the most complex circuits on the SPA and will ultimately determine the minimum latency of each SPA. Our research has thus led to designs for fast self-routing concentrators with logarithmic delays for use in terabit optical networks [29].

In the previous description, we assume that optical transmissions are pipelined between the boards. The backplane can also be designed so that the optical channels within the backplane are completely trasparent, i.e., an optical transmission from any board can be optically broadcasted to all other boards using an appropriate optical imaging system with fan-out. This variation has been called the *Transparent Hyperplane*. The same smart pixel designs can be used in the transparent version with minor variations.

Detailed operation. The operation of a smart pixel array programmed in the intelligent mode, which performs a relatively simple packet broadcasting scheme, is described. We describe an asynchronous backplane supporting several broadcast channels, where packets have any length and can arrive on an optical broadcast channel at any time. The pixels in a channel are always processing packets as they pass by, looking for packet headers with the appropriate address bits set. This processing is accomplished by comparing the address in the packet header with a

unique PCB address, using the address comparator cells in a channel. To speed up the address detection process, all address bits in a packet header can be compared with the PCB's unique address in parallel, using a straight-forward binary-tree type circuit (i.e., see [31]).

Unique PCB addresses are supplied from the MP to each smart pixel array and stored in an address-latch. In Fig. 8, the unique PCB address bits supplied from the MP are shown entering the top of the SPA. To conserve I/O pins, these address bits can be loaded bit-serially by the MP. When a channel recognizes its address in the header of a new packet, it asserts a *Receive Request* bit for that channel. The arbitration circuitry in the slice examines all the Receive-Request bits and generates the appropriate control signals for the extractors (i.e., concentrators) which causes the selected channels to enter the receiving state.

Each channel has its own channel control unit which stores the control signals which determine the state of the channel. The channel control units also contain the arbitration circuits which are used to generate control signals for the expanders or concentrators. Once a channel enters the receiving state, the packet can be delivered to the PCB over the electrical extractor channel I/O pins. Alternatively, the packet can be first delivered to an output queue on the smart pixel array where it is buffered, and then delivered to the PCB.

To configure the backplane, the appropriate control bits must be downloaded into the channel control units of each SPA. Each channel control unit requires typically 8 bits of control. The entire SPA also needs typically 8–16 bits to store its unique PCB address. The total number of control bits on the SPA is typically 1032 bits. Assuming that 1 electronic injector channel can be multiplexed to provide a byte of control per clock cycle, then approximately 129 clock cycles are required to completely reconfigure the SPA. At a clock rate of 250 Mhz, reconfiguration requires about 0.5 μs. Hence, the optical backplane can be completely reconfigured within a microsecond. In usual mode of operation the SPAs are reconfigured at initialization or each time an embedding is changed.

The above SPA design is conservative, and considerably more complex functions can be included within the smart pixel arrays. We are currently exploring considerably more intelligent arrays for optical backplanes, which include error and flow control, support for shared memory and synchronization, and other functions used in multiprocessor systems.

The message processors provide the interface and conversion between the data formats of the high speed optoelectronic smart pixel arrays and the application logic on the printed circuit boards. For reasonable speed and maximum flexibility, the message-processors can be implemented with Field Programmable Gate Arrays (FPGAs).

2.5. Scalalability to 2 and 3 Dimensions

Eventually expansion along one dimension will be limited by the increasing propagation delays over the length of the HyperPlane. A 1D HyperPlane with 256 PCBs or MCMs spaced 1 in. apart will be about 8 meters long and will have an end-to-end propagation delay of roughly 512 ns (assuming 2 ns latency per node).

The clock cycle time of a current electrical supercomputer is roughly 1 or 2 ns, and excessive propagation delays over a very large 1D optical backplane may cause unnecessary delays.

One solution is to *logically* extend the 1D HyperPlane into 2 or 3 dimensions, as shown in Fig. 10. A 2D logical structure consists of a 2D array of nodes, where each row and column supports 10–100s of optical broadcast channels (see Figs. 10a and 10b and the front face of Fig. 10c). Each row or column can be realized with an independent 1D HyperPlane, i.e., the rows and columns need not be directly optically connected.

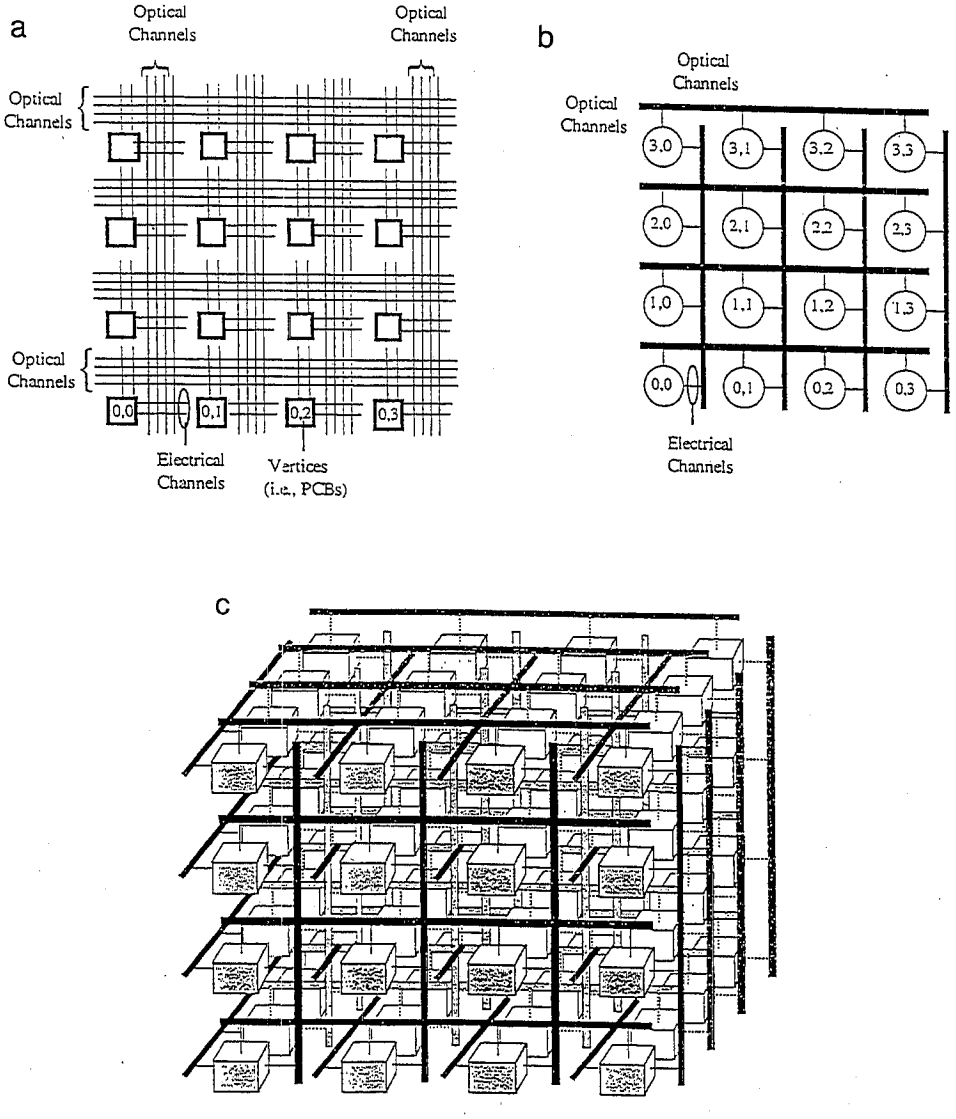

FIG. 10. (a) 2D HyperPlane. (b) Logical representation of a 2D HyperPlane based on HyperGraphs. (c) Logical representation of a 3D HyperPlane. Bold lines represent parallel broadcast channels.

This logical model can be *physically* realized in several ways. The 2D model can be physically implemented on a planar PCB substrate, using diffractive optics to interconnect optoelectronic MCMs or ICs on the PCB. This realization could potentially support 10s–100s of Tb/s of bandwidth on each PCB. Several planar PCB substrates can in principle be arranged in a linear array, as shown in Fig. 1, and optically interconnected in the 3rd dimension using optical transmission through transparent substrates (i.e., see [7, 14, 34] for a description of such technologies), yielding a compact 3-dimensional optical mesh-like structure, as shown in Fig. 10c. A 3D optical mesh-like structure will allow hundreds of nodes to be logically interconnected, as shown in Fig. 10c.

The 2D model can also be physically realized in a conventional cabinet using a straightforward approach. Several 1D backplanes as shown in Fig. 4a can be stacked on top of each other in a cabinet. Additional Optical Hardware Modules (OHMs) shown in Fig. 4b can then be used to interconnect the PCBs in each column. Thus, every row uses OHMs to implement a 1D HyperPlane in the X dimension, and every column uses additional OHMs to implement a 1D HyperPlane in the Y dimension. All optical connections in the X and Y dimensions are thus realized in a rigid plane at the back of the cabinet.

The logical model can also be implemented between physically distributed structures, using for example dense 2-dimensional optical image guides supporting thousands of optical bits (i.e., the type used in endoscope cables in the medical field, as suggested in [20]). Such flexible optical image guides can be used to interconnect a 2D array of OHMs at the back of a single cabinet, as described earlier, or to provide optical connections between several remote cabinets. Such distributed *Intelligent Optical Network* structures could potentially support 10–100s of Tb/s between boards and cabinets of electronics.

These higher dimensional optical mesh-like structures with multiple broadcast channels per row or column are essentially *Multi-Channel Meshes* or equivalently *HyperMeshes* [30]. They can be formally modeled as graph-theoretic *hypergraphs* and have some unique architectural attributes not present in conventional graph-based networks [30]. They also have powerful embedding capabilities. These structures will be analyzed in Section 3.

3. EMBEDDINGS IN THE MULTI-CHANNEL OPTICAL BACKPLANE

A general analysis of the reconfigurable multi-channel optical backplane is derived. The analysis models the key aspects of the backplane and its embedding capability, including the large aggregate optical bandwidth, the limited electrical "access" bandwidth, the minimum increment in which optical bandwidth can be allotted (equivalently, the bandwidth per optical channel), the topology to be embedded, the embedding strategy and the traffic model.

The analysis is based upon the classic *Optimal Capacity Assignment Problem* of general and arbitrary asynchronous M-channel N-node communication networks

proposed by Kleinrock [17] (also see [4] and [30]). In this model, the ith channel is represented as an $M/M/l$ queueing system with Poisson arrivals at a rate of λ_i packets per second and with a mean service time of $1/\mu C_i$ packets per second, where $1/\mu$ is the average packet size in bits, and C_i is the transmission bandwidth of the ith edge. With these assumptions the network is product-form and an *exact expression* for the expected delay is given by

$$T = \frac{\bar{k}\left(\sum_{i=1}^{M} \sqrt{\frac{\lambda_i}{\lambda}}\right)^2}{\mu C(1 - \bar{k} \cdot \rho)},$$

where ρ is the utilization factor, $\bar{k}$ is the average number of $M/M/1$ queues encountered by a packet, λ is the sum of all λ_i, and C is the sum of all the channel capacities C_i. The symbol ρ is the *utilization* of the network, defined as the average rate at which bits enter the network from external sources divided by C. Hence, $\rho = 1$ denotes a fully loaded network and $\rho > 1$ denotes a network where the demand placed upon the network exceeds its capacity to accept or deliver, and hence the queueing delays approach infinity. In all vertex-symmetric networks, the second term in the numerator yields the number of channels in the network M, and the term (C/M) equals the bandwidth of any edge denoted C_{edge}, hence the expected delay can be simplified to yield [30]

$$T = \frac{\bar{k} \cdot M}{\mu C(1 - \bar{k} \cdot \rho)} \quad \text{or} \quad T = \frac{\bar{k}}{\mu C_{\text{edge}}(1 - \bar{k}\rho)}.$$

We note that as $\rho \to 0$ the expected delay reduces to $k/\mu C_{\text{edge}}$, equal to the average number of queues encountered by a packet times the expected service time per packet. Hence, at light loads the delay of a packet is simply its transmission time times the number of "hops" it takes.

The Optimal Capacity Assignment Problem yields the exact queueing delays given its assumptions. It can be used to evaluate the performance of arbitrary topologies, including electrical or optical topologies, and it is particularly useful to identify the peak usable bandwidth capacity of a topology, i.e., the load above which its delays become unbounded.

Assume a *random-uniform* traffic model, where each node is equally likely to send a message to every other node in a fixed period of time. With this traffic model $\bar{k}$ is easily determined for a given topology. (A traffic model with locality is easily incorporated into the analysis and only affects $\bar{k}$.) In the following analysis, C and C_{edge} for each topology to be embedded are determined, and are a function of the cut-width of the embedded topology, the embedding strategy (since there are many ways to embed a given topology), and the number of optical and electrical channels assigned to each embedded edge.

When a graph is embedded into the optical backplane, the total delay of a packet is given by the sum of the expected queueing delays and the expected propagation delays. The proposed analysis yields the expected queueing time only. The expected

propagation delay is much easier to determine and is on average given by the time needed for an optical signal to traverse half the backplane.

The optical backplane consists of two optical ring-like datapaths in the *upstream* and *downstream* directions, as shown in Figs. 4a and 5. These datapaths will be called *streams*. Let each PCB have two conservative smart pixel arrays, as described in section 2, in each stream. Therefore, each stream supports 64 optical channels, and each PCB (or "node") has 8 electrical channels to and from each stream, where all channels support 16 Gb/s each. Each PCB can therefore inject and extract 128 Gb/s of data into each stream. The aggregate electrical bandwidth can be defined as the peak bandwidth which can be injected into the backplane by all PCBs, or 4 Tb/s. The aggregate optical bandwidth can be defined as the peak optical bandwidth which can be carried by the backplane and is 2 Tb/s, equivalent to the bandwidth of 128 optical channels. Hence, in this design example the maximum amount of data which can be moved across the backplane is upper bounded by 2 Tb/s. However, these parameters can have a range of values and are determined by the system designer, as described in Section 2.

During the embedding analysis, it is often convenient to consider all electrical and optical channels as "bi-directional" and to consider only one stream (rather than two streams in separate directions). The "cut-width" of an embedding into the HyperPlane is defined as the number of embedded edges which exist between any two neighboring nodes, in any one stream. The following technology parameters are used:

- N = number of nodes (PCBs or MCMs) in the optical backplane = 16
- C_{opt} = number of optical channels in a backplane stream = 64
- C_{ele} = number of electrical access channels per node per stream = 8
- C = sum of all channel (edge) capacities of the embedded topology
- C_{edge} = bandwidth of any edge of the embedded topology
- $1/\mu$ = expected length of a packet in bits = 512

3.1. Single Ring or Bus (1-Dimensional Array)

The graph model of a single ring, or equivalently, a "1D mesh" or "linear array" with wrap-around, is shown in Fig. 11a. Each node in the ring has an edge to the nearest neighbor in each direction. The embedding of a broadcast bus spanning all nodes is similar and in this paper no distinction will be made between a ring and a bus. (We assume that both the ring and bus topologies are "pipelined," i.e., packets of data are latched at each smart pixel array as they travel down the backplane.) An embedding of a ring into the multichannel optical backplane is shown in Fig. 11b. The maximum cut-width in any stream is 1; i.e., at most one edge exists between any two neighboring nodes. Hence, each embedded edge can be allocated up to $C_{opt}/1 = 64$ optical channels. However, each edge can only be "driven" by 8 electrical channels since each node has only has 8 electrical injection channels per stream. Hence, each embedded edge of the ring consists of 8 optical channels operating in parallel, for an effective bandwidth of 128 Gb/s. In other words, the optical ring (or pipelined bus) is 256 bits wide and is clocked at 500 Mhz.

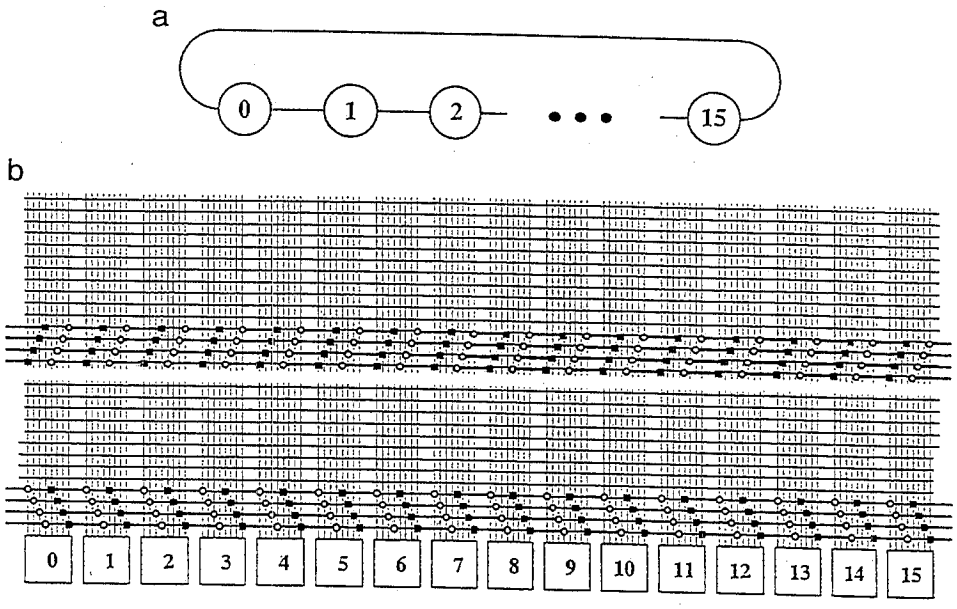

FIG. 11. (a) 1D ring or bus topology. (b) Embedding using circular Hyperplane.

For a single ring with N nodes, the average distance is $N/4$, and the effective bandwidth per edge $C_{\text{edge}} = 128$ Gb/s. According to Kleinrock's model, $C = 2NC_{\text{edge}} = 4$ Tb/s, and the expected queueing time is given by

$$T = \frac{(1/\mu)\cdot \bar{k}}{C_{\text{edge}}(1-\bar{k}\rho)}\ \text{s} = \frac{512\cdot(N/4)}{128\cdot(1-(N/4)\cdot\rho)}\ \text{ns} = \frac{16}{(1-4\rho)}\ \text{ns}. \tag{1}$$

In this embedding, the optical backplane is moderately utilized, since the single ring or bus with an edge bandwidth of 128 Gb/s is insufficient to fully utilize the optical bandwidth available in each stream of the optical backplane. Equivalently, the embedding of a ring with 256 optical bits per edge does not fully utilize the 4096 optical bits available in the optical backplane. The queueing delays become unbounded as $4\rho \to 1$, i.e., as the offered load approaches $C/4 = 1$ Tb/s.

3.2. 2-Dimensional Meshes

A regular embedding of the 2D mesh (with wrap-around edges) is shown in Fig. 12a. The 2D d^n mesh ($d=4$, $n=2$) can be viewed as d parallel rings in each of n dimensions [8, 28]. Let each ring in dimension i ($0 \leqslant i \leqslant n-1$) interconnect d nodes. The rings in dimension 0 together have a cut-width of 2 and can be embedded using two optical channels after they are partitioned (i.e., the rings are nonoverlapping once they are embedded into the backplane, and the two optical channels used to embed one ring can be partitioned to embed all rings). Each ring in dimension 1 contributes 2 to the cut-width, as shown in Fig. 12b. Hence, the d rings in dimension 1 require $2d$ optical channels to be embedded. Hence, the maximum cut-width of the 2D mesh is $2+2d$; i.e., $2+2d$ edges exist simultaneously at some point in any stream of the backplane. For the mesh d^n with $d=4$, $n=2$, the optical backplane supports up to 10 embedded edges in each stream. Therefore, each

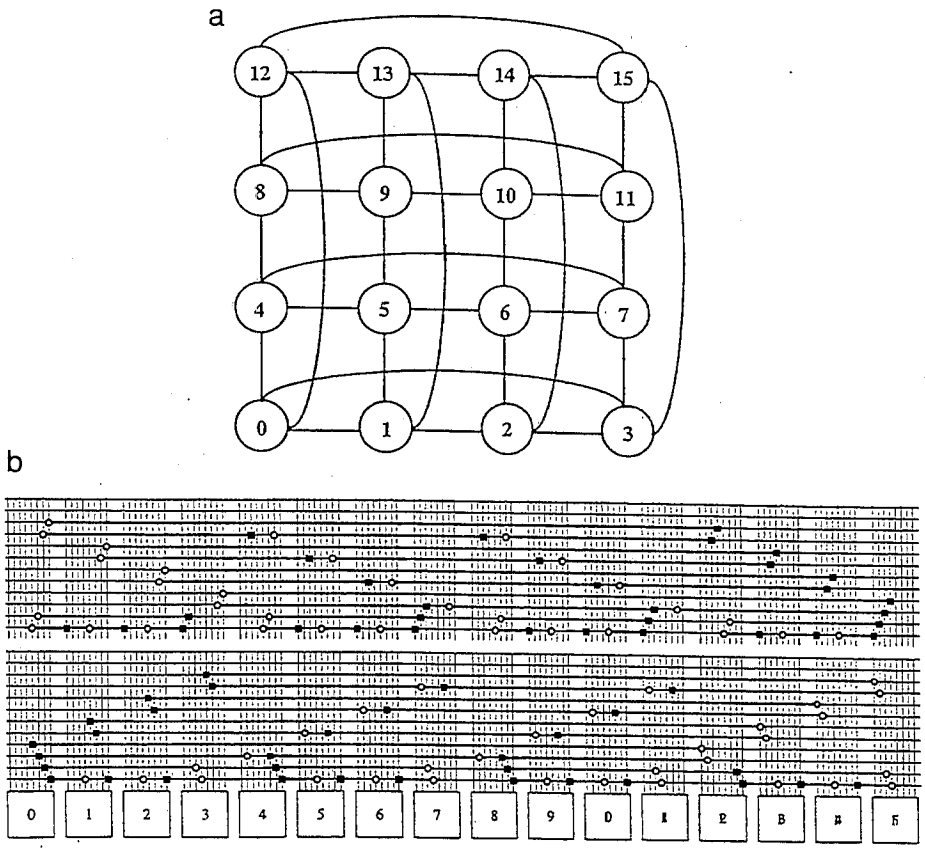

FIG. 12. (a) 2D mesh topology. (b) Embedding using circular Hyperplane.

embedded edge can be allocated up to $\lfloor C_{opt}/10 \rfloor = 6$ optical channels. Each mesh node has degree 4 and must therefore support up to 4 edges per stream. Since each node has 8 electrical injector channels per stream, then each embedded edge can be driven by 2 electrical channels. The effective bandwidth for an embedded edge is the minimum of the electrical and optical bandwidths of the edge, i.e., $C_{edge} = 2$ channels * 16 Gb/s = 32 Gb/s in this case. Equivalently, each edge consists of 64 optical bits clocked at 500 Mhz.

For a d^n mesh with $N = 16$ nodes, the average distance is $nd/4 = 2$, $C_{edge} =$ 32 Gb/s, $C = 4NC_{edge} = 2$ Tb/s, and the expected queueing time is given by

$$T = \frac{(1/\mu)\cdot \bar{k}}{C_{edge}(1-\bar{k}\rho)}\ \text{s} = \frac{512\cdot(nd/4)}{32\cdot(1-(nd/4)\cdot\rho)}\ \text{ns} = \frac{32}{(1-2\rho)}\ \text{ns}.$$

In this embedding, the optical backplane is moderately utilized since the 2D ring does not fully utilize the full optical bandwidth available in each stream of the optical backplane. Equivalently, the embedding of a 2D mesh with 64 optical bits per edge and a bisection bandwidth of 10 edges utilizes only 640 optical bits per stream, and does not fully utilize the 4096 optical bits available in the optical backplane. The queueing delays become unbounded as $2\rho \to 1$; i.e., the offered load approaches $C/2 = 1$ Tb/s.

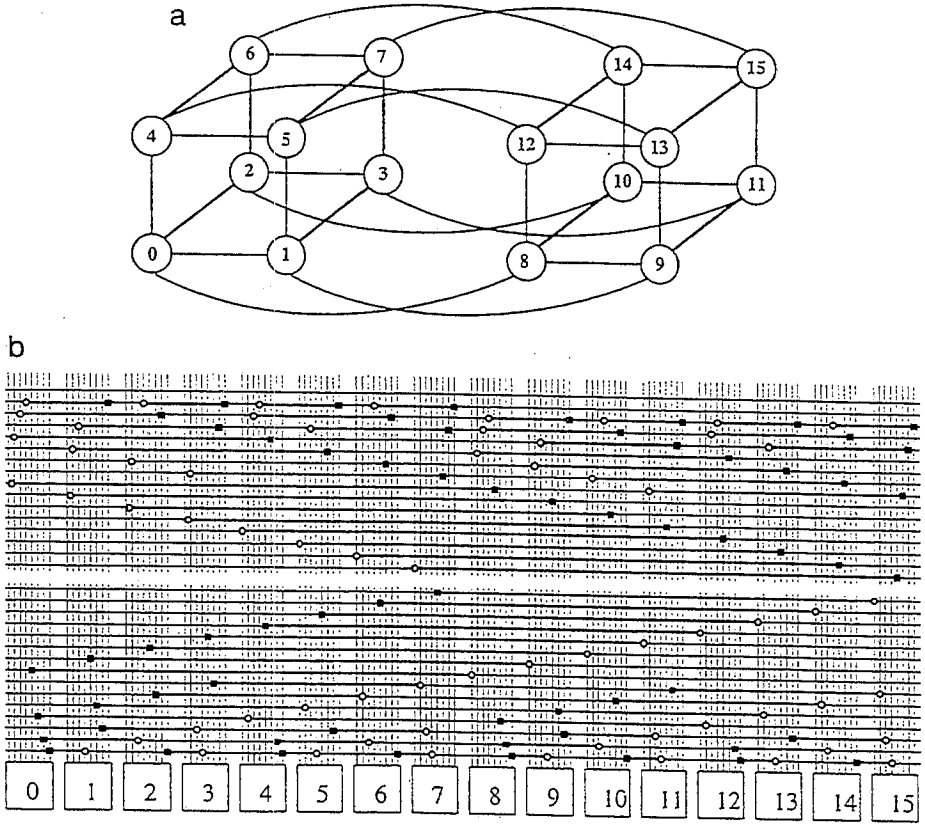

FIG. 13. (a) Binary hypercube topology. (b) Embedding using circular Hyperplane.

3.3. Binary and Higher Radix Hypercubes

A regular embedding of the binary hypercube is shown in Fig. 13a. The cut-width of an N node hypercube is N-1. For a binary hypercube, we summarize the key parameters. The average distance is $\bar{k} = \log_2 N/2$, the number of edges is $N \log_2 N$, the bandwidth per edge is $C_{\text{edge}} = 32$ Gb/s (the limitation is the 8 electrical injector channels needed to support 4 outgoing edges in each stream). Thus $C = 4NC_{\text{edge}} = 2$ Tb/s. Hence, the expected queueing time is given by

$$T = \frac{(1/\mu)\cdot\bar{k}}{C_{\text{edge}}(1-\bar{k}\rho)}\ \text{s} = \frac{512\cdot(\log_2 N/2)}{32\cdot(1-(\log_2 N/2)\cdot\rho)}\ \text{ns} = \frac{32}{(1-2\rho)}\ \text{ns}.$$

The hypercube delay is equal to the 2D mesh delay (Section 3.2), which is expected since these networks are topologically equivalent for this size. The results of the generalized higher radix hypercubes described in [4] have been derived with the same methodology and are also shown in the figures.

3.4. One-Dimensional Multiple-Bus/Multiple-Ring System

To fully utilize the optical bandwidth, one may embed multiple rings or broadcast buses into the backplane (we make no distinction between a pipelined bus and a

slotted ring). Given 64 optical channels in the backplane, one may embed 8 bidirectional rings (or buses). Each ring or bus will contain 8 optical channels operating in parallel, as shown in Fig. 14. Each PCB will thus require 8 electrical injector channels to feed any optical ring/bus. In this embedding, each PCB can broadcast on any one ring/bus out of 8 available rings/busses, where each ring/bus is 256 bits wide and is clocked at 500 Mhz. Each PCB can receive packets from all 8 rings/buses, as the smart pixel arrays can act as packet filters. This topology is essentially a "broadcast-and-select" topology.

For this multi-channel bus or ring network, the analytic model is similar to the single ring in Section 3.1. We analyze any one ring/bus, since all rings/buses are statistically identical in this model. However, compared to the single ring in

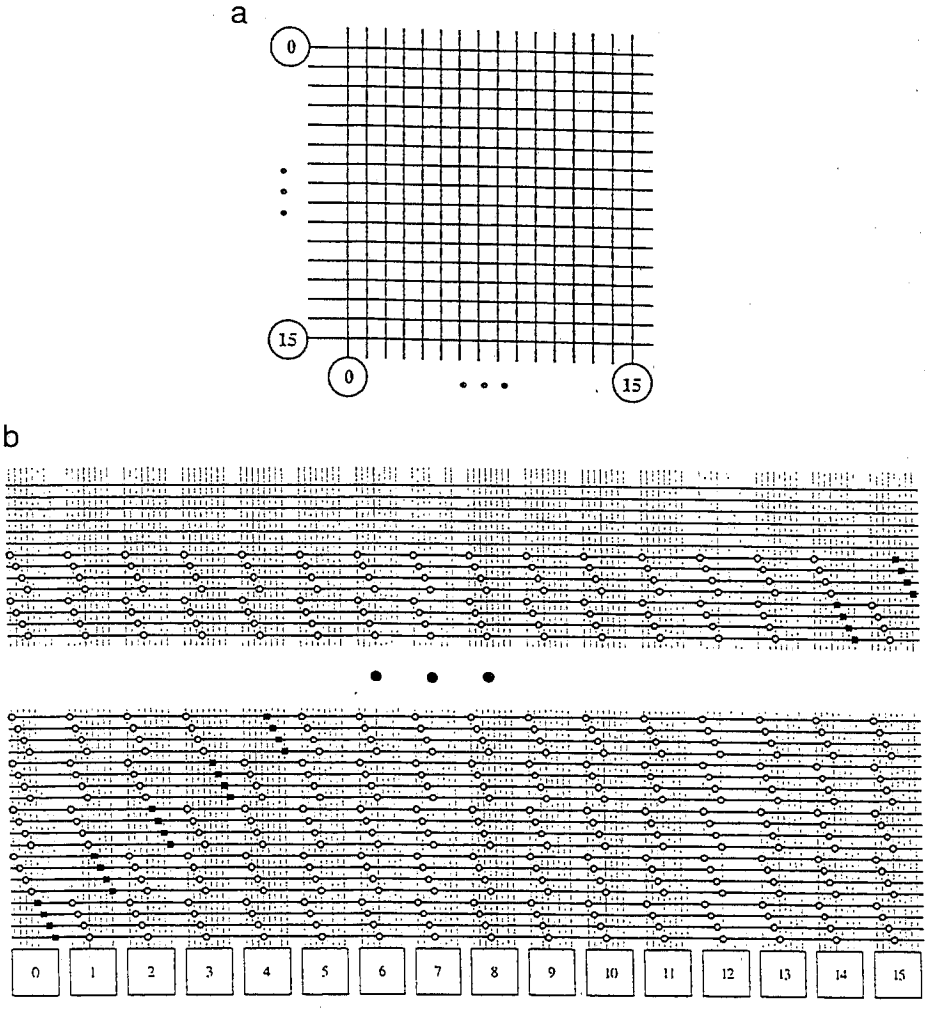

FIG. 14. (a) Multichannel "broadcast-and-select" topology (multiple rings/buses). (b) Embedding using circular Hyperplane. All optical bits are fully utilized.

Section 3.1, ρ is divided by 8 since the traffic generated is now distributed over 8 rings/buses. Thus, $C_{\text{edge}} = 128$ Gb/s (the limitation is the 8 optical channels per ring/bus), $C = 2NC_{\text{edge}} = 4$ Tb/s, and the expected queueing time is given by

$$T = \frac{(1/\mu)\cdot \bar{k}}{C_{\text{edge}}(1-\bar{k}\rho)}\ \text{s} = \frac{512\cdot(N/4)}{128\cdot(1-(N/4)\cdot\rho/8)}\ \text{ns} = \frac{16}{(1-2\rho)}.$$

The queueing delays over any one ring/bus never becomes unbounded, since the total load is distributed over multiple buses such that every bus is moderately loaded. Hence, this topology utilizes the full 2 Tb/s bandwidth of the backplane, i.e., it utilizes all 4096 optical bits clocked at 500 Mhz. At full load (2 Tb/s), the queueing delay is a modest 20 ns.

3.5. Embedding Comparisons

Figure 15a illustrates the queueing delay versus offered load for various topologies embedded into the 1D HyperPlane, using the conservative SPA design from Section 2. Three basic families of topologies are embedded: (i) rings and meshes, (ii) hypercubes, and (iii) multiple bus-based networks such as crossbars and hypermeshes. Packets have a mean length of 512 bits. In Fig. 15a, the ring, the mesh and the hypercubes all tend to moderately utilize the bandwidth of the optical backplane, and all exhibit unbounded delays as the load approaches 1 Tb/s. The multiple bus/ring network (called the 1D HyperMesh in Fig. 15a) is the most efficient topology, since it utilizes the full 2 Tb/s bandwidth of the optical backplane. In this topology, the backplane supports 8 buses/rings, each 256 bits wide clocked at 500 Mhz. The average delay is in the neighborhood of 20 ns, which is suitable for large-scale shared memory multiprocessing. The analysis does not include propagation delays. Assuming it takes 4 ns to traverse each PCB, the average propagation delay through the backplane is equal to the time to traverse 8 PCBs, or about 32 ns on average. This propagation delay is deterministic, and will represent an additional delay above the queueing delays shown in Figs. 15 and 16.

Figure 15b illustrates queueing delay versus offered load for various topologies embedded into the 2D HyperPlane, using the conservative SPA design from Section 2. The 2D HyperMesh topology, essentially a 2D mesh with multiple broadcast-based rings/buses (channels) in each row and column, is the most efficient embedding since it fully utilizes the optical bandwidth of each row and column. In this topology, each row or column supports 8 rings/buses, each 256 bits wide clocked at 500 Mhz. These results are consistent with [30], where it was shown that the 2D mesh-like structure with multiple broadcast channels in each row or column is ideal at exploiting the bandwidth advantage of fiber optic networks.

In Fig. 16 we consider the capability of an optical backplane a few years into the future. Let each advanced SPA support 256 optical channels and 32 electrical access channels at 16 Gb/s each, i.e., 8 times the capacity of the conservative SPAs used in Fig. 15. The packets have a mean length of 4096 bits. The 1D backplane supports

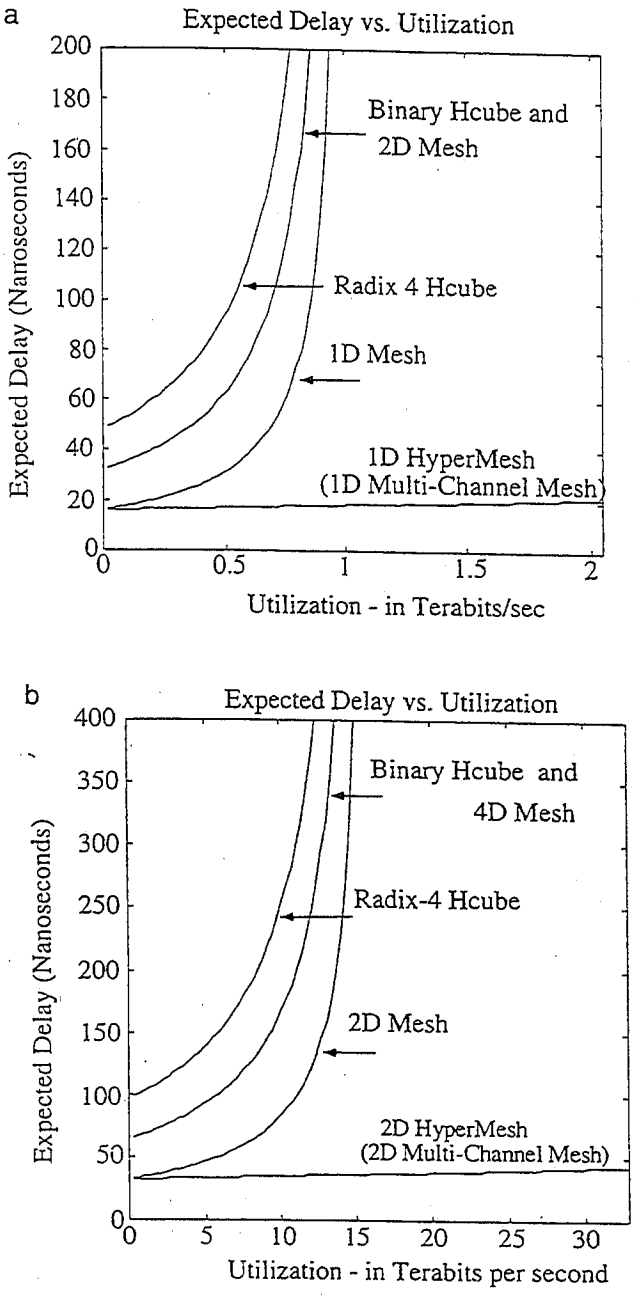

FIG. 15. Performance of 1D and 2D optical backplanes (conservative SPA). (a) 1D 16 × 1 HyperPlane. (b) 2D 16 × 16 HyperPlane.

512 optical channels in each stream, and the backplane has an aggregate optical bandwidth of 16 Tb/s. The one-dimensional backplane with 16 PCBs can be logically extended to a 2-dimensional structure with 256 nodes, where each row or column supports 512 optical broadcast channels in each stream. The aggregate optical bandwidth of the 2D structure is 256 Tb/s. (The use of WDM could potentially increase the bandwidth by another order of magnitude.)

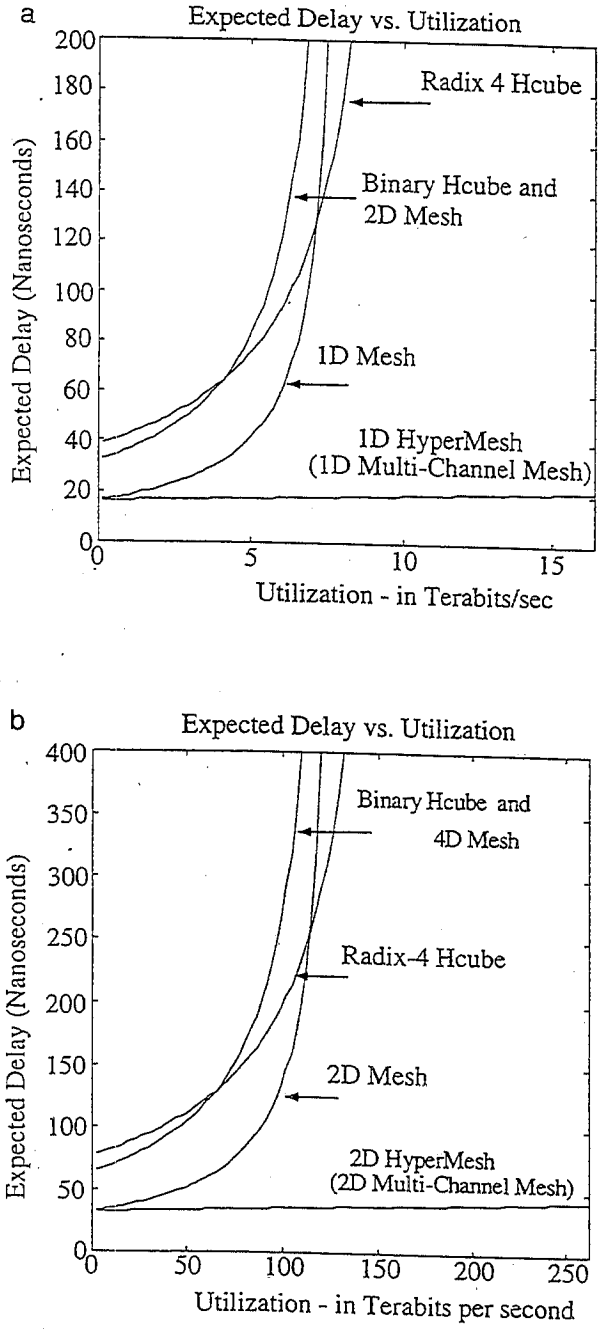

FIG. 16. Performance of 1D and 2D optical backplanes (advanced SPA). (a) 1D 16 × 1 HyperPlane. (b) 2D 16 × 16 HyperPlane.

Figure 16 illustrates the queueing delays versus load for various topologies embedded into 1D and 2D HyperPlanes. Referring to Fig. 16a, in the 1D backplane the multiple-ring/bus topology (the 1D Hypermesh) is the most efficient, since it fully utilizes the 16 Tb/s of bandwidth. Referring to Fig. 16b, once again the 2D

mesh-like topology with multiple broadcast-based rings/buses in each row or column (the 2D Hypermesh) is the most efficient embedding, since it fully utilizes the optical bandwidth of each row and column, and can support offered loads of up to 256 Tb/s. The expected queueing delay is between 20 and 50 ns, which is suitable for shared memory multiprocessors.

The relative performances of the three families of topologies (meshes, hypercubes, and multiple ring/bus-based networks such as HyperMeshes) tend to remain constant over a wide range of parameter values. Furthermore, as the ratio of optical to electrical bandwidth increases, the ring/bus based networks tend to become better. To further illustrate the capabilities of these optical structures, embeddings for other complex graphs such as Stars are shown in [23].

The ring/bus-based topology is well suited for supporting advanced communication protocols in the optical backplane. The smart pixel arrays can be programmed to act as packet filters, which extract packets which meet some criterion for extraction. Since every channel spans all PCBs, every channel therefore naturally supports multicasting and broadcasting.

3.6. Applications—Shared Memory Multiprocessors and NOWs

The optical backplane also has particular appeal for existing distributed shared memory multiprocessor systems. In this model, all processors maintain a local fraction of the distributed shared memory, and each processor also maintains a cache (i.e., a snoopy cache) of globally shared variables. The processors maintain cache consistency by broadcasting changes of shared variables to all other processors. Every cache controller monitors all broadcasts throughout the system and updates its own local cache. This model is generally considered relatively unscalable beyond perhaps a hundred processors, due to the inability of an electrical backplane to support the global broadcasts. The concept of providing each PCB in a backplane access to a lightly loaded broadcast bus, with 256 bits clocked at 500 Mhz, can potentially facilitate large-scale shared memory multiprocessing. Hence, an intelligent optical backplane can significantly extend the scalability of shared memory multiprocessors, by providing Terabits of low latency bandwidth.

The optical backplane can also be used to interconnect workstations in a Network-of-Workstations (NOW). Current microprocessors require $\sim$10 million transistors and generate I/O bandwidth at the rate of $\sim$1–5 Gb/s (assuming a 10% miss rate for the on-chip memory cache). In a decade, single chip microprocessors are expected to utilize up to 100 million gates or more, and may thus contain the equivalent of up to 10 of today's high-performance microprocessors on the same die. Hence, in a decade the I/O bandwidth of each workstation may be in the 10–50 Gb/s range. Consider a NOW with 64 high-performance workstations, each with a high-bandwidth connection to a central backplane with 16 PCBs. Each PCB would support the fiber ribbons of four workstations. Each connection may support up to 32 Gb/s bandwidth to and from the backplane, and each PCB would thus generate up to 128 Gb/s of bandwidth for the backplane. The backplane must support up to 2 Tb/s of bandwidth, and it is unclear if electrical backplanes can be

scaled to these capacities and beyond. The proposed intelligent optical backplane could for example support a reserved 32 Gb/s broadcast bus for each workstation. Intelligent optical backplanes thus provides one potential means for allowing clusters of workstations to scale to support terabits of low-latency bandwidth.

4. CONCLUSIONS

Within the computing community, there is a growing awareness that optical interconnects may begin to appear in the board-to-board packaging hierarchy within a few product generations. This paper argues that optics can potentially open up a new frontier of computing machines with massive connectivity not previously possible, and has described one possible way to exploit dense bit-parallel optical technology, in the form of a terabit Intelligent Optical Backplane. A one dimensional backplane can potentially provide between 1 and 100 Tb/s of low-latency bandwidth distributed over 16 or 32 PCBs (or MCMs). The intelligent optical backplane can also provide built-in hardware support for common communication primitives used in shared memory multiprocessing, including multicasting, broadcasting, acknowledgment, and error and flow control. The merging of CMOS logic with optical I/O allows for a level of "intelligence" not possible with other optical technologies, and in the future smart pixel arrays may include support for shared memory, synchronization, and other operations used in multiprocessors.

The architecture also scales to higher dimensions. It was shown that 2- and 3-dimensional optical mesh-like structures, with hundreds of high bandwidth reconfigurable optical broadcast channels in each row or column, are feasible. Such architectures can potentially provide between 10 and 1000 Tb/s of low-latency optical bandwidth. In the future, one of the challenges facing architects may be defining new multiprocessing architectures which exploit the bandwidth of optics.

An Intelligent Optical Backplane is under development at McGill University in Canada, funded in part by the Canadian Institute for Telecommunications Research (CITR), a member of the federal Networks of Centers of Excellence (NCE) program in Canada. The research program is multi-disciplinary and includes the collaboration of many institutes and researchers, spanning device technology, micro-optics, packaging and architectures. The first generation backplane demonstrator was designed to be relatively conservative, consisting of a few channels. The optical design and testing aspects are described in [24]. The VLSI layout and testing of the first generation smart pixel arrays are described in [27]. Extensible Optical Hardware Modules are described in [11, 26]. Colleagues at the University of Colorado Rave considered buffering and token passing in the optical backplane in [37, 38]. Advanced forms of intelligent backplane processing in support of shared memory multiprocessing are described in [31] and [32].

A second generation intelligent optical backplane demonstrator will build upon our experiences with the first and will be more aggressive. The optical backplane is expected to include 256 optical bits clocked at a few hundred Mb/s each and will support the NUMAchine shared memory multiprocessor developed at the University of Toronto [39].

ACKNOWLEDGMENTS

We thank the reviewers for their comments. This research was supported by NSERC Canada Grant OGP 0121601, by the Canadian Institute for Telecommunications Research (CITR) under research Grants 1993-3-4 through 1998-3-4, and the Hudson Moore Jr. Professor of Engineering at the University of Colorado at Boulder. The support of the Canadian Microelectronics Corporation (CMC) is acknowledged. Thanks to the graduate students of the Microelectronics and Computer System Laboratory at McGill, and to the professors and students of the Photonic Systems Laboratory at McGill, including Professor David Plant and Professor Andrew Kirk, and Professor Frank Tooley and Dr. Brian Robertson of Heriot-Watt University.

REFERENCES

1. S. Abraham and K. Padhmanabhan, Performance of multicomputer networks under pin-out constraints, *J. Parallel Distrib. Comput.* **12** (1992), 237–248.
2. A. Barak and E. Schenfeld, Embedding classical communication topologies in the scalable OPAM architecture, *IEEE TPDS* 7, No. 9 (1996), 962–978.
3. J. Bhasker and S. Sahni, Optimal linear arrangement of circuit components, *J. VLSI Computer Systems* (1987), 87–109.
4. L. N. Bhuyan and D. P. Agrawal, Generalized hypercube and hyperbus structures for a computer network, *IEEE Trans. Comput.* **C-33**, No. 4 (1984), 323–333.
5. J. A. Bondy and U. S. R. Murphy, "Graph Theory with Applications," North-Holland, Amsterdam.
6. T. J. Cloonan, Comparative study of optical and electronic interconnection technologies for large asynchronous transfer mode packet switching applications, *Optical Eng.* **33**, No. 5 (1994), 1512–1523.
7. C. Camperi-Ginestet, B. Buchanan, N. M. Jokerst, and M. A. Brooke, Bi-directional communication through stacked silicon circuitry using integrated thin film InP-based emitters and detectors, *in* "Proc. Conf. on Lasers and Electro-Optics, Baltimore, June 1995."
8. W. J. Dally, Performance analysis of k-ary n-cube interconnection networks, *IEEE Trans. Computers* **39**, No. 6 (1990), 775–785.
9. W. J. Dally and J. Poulton, Transmitter equalization for 4-Gbps signaling, *IEEE Micro.* **17**, No. 1 (1997), 48–56.
10. P. Dowd, Wavelength division multiple access channel hypercube processor interconnection, *IEEE Trans. Computers*, **41**, No. 10 (1992), 1223–1241.
11. D. J. Goodwill and H. S. Hinton, Compact and extensible optical interconnect module for a free-space optical backplane, *in* "SPIE Proceedings Vol. 2692, Optical Interconnects in Broadband Switching Architectures" (T. J. Cloonan), Jan. 1996.
12. K. Goossen, J. A. Walker, L. A. D'Asaro, S. P. Hui, B. Tseng, R. Leibenguth, D. Kossives, D. D. Bacon, D. Dahringer, L. M. F. Chirovsky, A. L. Lentine, and D. A. B. Miller, GaAs MQW modulators integrated with silicon CMOS, *IEEE PLT* 7 (1995), 360–362.
13. Z. Guo, R. Melhem, R. Hall, D. Chiarulli, and S. Levitan, Pipelined communications in optically interconnected arrays, *J. Parallel Distrib. Comput.* **12** (1991), 269–282.
14. K. Hamanaka, Optical bus interconnection using selfoc lenses, *Optics Lett.* **16**, No. 16 (1991), 1222–1224.
15. J. L. Hennessy and D. A. Patterson, "Computer Architecture: A Quantitative Approach," Morgan Kauffman, San Mateo, CA, 1995.
16. H. S. Hinton, T. J. Cloonan, F. A. P. Tooley, and F. McCormick, Free space digital optical systems, *Proc. IEEE* **82**, No. 11 (1994), 1632–1649.
17. L. Kleinrock, "Queueing Systems," II, Wiley, New York, 1975.

18. A. V. Krishnamoorthy and D. A. B. Miller, Scaling optoelectronic-VLSI circuits into the 21st century: A technology roadmap, *IEEE J. Selected Topics Quantum Electronics* **2**, No. 1 (1996), 55–76, April 1996.

19. F. T. Leighton, "Parallel Algorithms and Architectures: Arrays, Trees and Hypercubes," Morgan Kauffman, San Mateo, CA, 1992.

20. Y. Li *et al.*, Applications of fiber image guides to bit-parallel optical interconnections, *in* "Proc. Int. Conf. Optical Computing, March 13–15, 1995."

21. S. Matsuo, T. Nakahara, Y. Kohama, Y. Ohiso, S. Fukushima, and T. Kurokawa, Monolithically integrated photonic switching device using an MSM PD, MESFET's, and a VCSEL, *IEEE Photonics Technol. Lett.* 7, No. 10 (1995), 1165–1167.

22. R. A. Nordin *et al.*, A systems perspective on digital interconnection technology, *J. Lightwave Technol.* **10**, No. 6 (1992), 811–827.

23. S. T. Obenaus and T. H. Szymanski, Embedding star graphs into optical meshes without bends, *J. Parallel Distrib. Comput.* **44**, No. 2 (1997), 97–106. [Ph.D. thesis, in prep.]

24. D. V. Plant, B. Robertson, H. S. Hinton, M. H. Ayliffe, G. C. Boisset, D. J. Goodwill, D. Kabal, R. Iyer, Y. S. Liu, D. R. Rolston, M. Venditti, T. H. Szymanski, W. M. Robertson, and M. R. Taghizadeh, Optical, optomechanical and optoelectronic design and operational testing of a multistage optical backplane demonstration system, *MPPOI* (1996), 306–312.

25. I. Redmond and E. Schenfeld, A distributed reconfigurable free-space optical interconnection network for massively parallel processing architectures, *in* "Proc. Int. Conf. Optical Computing, Edinburgh, 1994," pp. 215–218, Institute of Physics Publishing, 1995.

26. B. Robertson, Design of a compact alignment tolerant optical interconnect for photonic backplane applications, *MPPOI-97* (1997), 68–77.

27. D. R. Rolston, D. V. Plant, T. H. Szymanski, H. S. Hinton, M. H. Ayliffe, D. N. Kabal, A. V. Krishnamoorthy, K. W. Goosen, J. A. Walker, B. Tseng, S. P. Hui, J. C. Cunningham, and W. Y. Jan, A hybrid-SEED smart pixel array for a four-stage intelligent optical backplane demonstrator, *J. Quantum Electronics* (1996), 97–105.

28. I. Scherson, Orthogonal graphs for a class of interconnection networks, *IEEE Trans. Parallel Distrib. Systems* **2**, No. 1 (1991), 3–19.

29. B. Supmonchai and T. H. Szymanski, High speed VLSI concentrators for terabit intelligent optical backplanes, *in* "Proc. Int. Conf. Optical Computing, 1998." [and Ph.D. thesis, in prep.]

30. T. H. Szymanski, Hypermeshes-optical interconnection networks for parallel computing, *J. Parallel Distrib. Comput.* **26** (1995), 1–23.

31. T. H. Szymanski and H. S. Hinton, Reconfigurable intelligent optical backplane for parallel computing and communications, *Appl. Opt.* (1996), 1253–1268.

32. T. H. Szymanski and H. S. Hinton, "Optoelectronic Smart Pixel Arrays for a Reconfigurable Intelligent Optical Interconnect," U.S. Patent Application.

33. Semiconductor Industry Association, "The National Technology Roadmap for Semiconductors," SIA, San Jose, 1994.

34. D. S. Wills, W. S. Lucy, C. Camperi-Ginestet, B. Buchanan, S. Wilkinson, M. Lee, N. M. Jokerst, and M. Brooke, A fine grain high throughput architecture using through wafer interconnect, *J. Lightwave Technol.* **13** (1995), 1085–1092.

35. T. K. Woodward, A. L. Lentine, K. W. Goosen, J. A. Walker, B. T. Tseng, S. P. Hui, J. Lothian, and R. E. Leibenguth, Demultiplexing 2.48 Gb/s optical signals with a CMOS receiver array based on clocked-sense amplifier, *IEEE Photonics Technol. Lett.* **9**, No. 8 (1997), 1146–1148.

36. B. Zerrouk, A. Greiner, V. Reibaldi, F. Potter, A. Derieux, R. Marbot, and R. Nezamzadeh, The HIC high speed link technology and associated router, *Real-Time Magazine* **96**, No. 3, 73–77.

37. D. J. Goodwill, K. D. Davenport, and H. S. Hinton, An ATM-based intelligent optical backplane using CMOS/SEED smart pixel arrays and free-space optical interconnection modules, *IEEE J. Selected Topics in Quantum Electronics*, April 96.

38. B. E. Angliss and H. S. Hinton, A token based dynamically reconfigurable optical backplane, *in* "OSA 1997 Spring Topic Meeting on Optics in Computing," 1997.

39. Z. Vranesic, S. Brown, M. Stumm, S. Caranci, A. Grbic, R. and Grindley, *et al.*, "The NUMAchine Multiprocessor," CSRI Technical Report, University of Toronto.

TED H. SZYMANSKI received a Bachelor's degree in engineering science and a Ph.D. in electrical engineering from the University of Toronto. He is currently an associate professor at McMaster University. He has served as an associate professor and the Director of the Microelectronics and Computer Systems (MACS) Laboratory at McGill University in Montreal, and since 1993 as a project leader in the Canadian Institute for Telecommunications Research. An "Intelligent Optical Backplane" architecture developed by this project will be demonstrated in Canada, as part of a ten year multi-institutional research program. He has presented several invited talks and papers at international conferences, and several of his papers on switching systems have been reprinted in *IEEE* textbooks. He is active professionally, and has been on the program committees for the 1999, 1998, and 1997 *Workshops on Optics in Computer Science*, the 1998 and 1997 *International Conferences on Massively Parallel Processing Using Optical Interconnects*, the 1998 *International Conference on Optical Computing*, the 1997 *Innovative Systems on Silicon Conference*, the 1995 *Workshop on High-Speed Network Computing*, and the 1998, 1995, and 1994 *Canadian Conferences on Programmable Logic Devices*. He has consulted for several companies, including Spar Space Systems and the Strategic Microelectronics Consortium of Canada. His personal interests include snowboarding, and his research interests include optical networks, telecommunication and computing systems, and performance analysis. He is a member of the *IEEE* Computer and Communications societies.

H. SCOTT HINTON was born in Salt Lake City, Utah in 1951. He received a B.S.E.E. in 1981 at Brigham Young University and a M.S.E.E. at Purdue University in 1982. In 1981 he joined AT&T Bell Laboratories in Naperville, IL and eventually became the Head of the Photonic Switching Department in 1989. From 19921994 he was the BNR–NT/NSERC Chair in Photonic Systems at McGill University. In 1994 he accepted a position at the University of Colorado at Boulder as the Hudson Moore Jr. Professor of Engineering. Professor Hinton has been active in optoelectronic systems research where he has published over 30 articles, co-edited 4 books, contributed to 12 books, presented over 80 conference papers, and has been awarded 12 patents. His current interests are in systems based on smart pixel arrays and free-space optical interconnects. He is a fellow of both the IEEE and OSA.

Reprinted from Journal of Parallel and Distributed Computing
55, 1-31 (1998)